# People's Paper[1]

HK-230965-15
Reading the First Books: Multilingual, Early-Modern OCR for Primeros Libros
Project Directors: Sergio Romero, Laura Mandell, Anton DuPlessis
University of Texas at Austin and Texas A&M University
NEH Office of Digital Humanities Digital Implementation Grant
August 2015 - December 2017
Date Report Submitted: _____

# Table of Contents



## Appendices



[1] We follow the recommendation of @ccp_org (Colored Conventions) in replacing the historical term for this document with more descriptive and specific language.

# People's Paper

## Reading the First Books: Multilingual, Early Modern OCR for Primeros Libros

In the summer of 2015, the National Endowment of the Humanities Office of Digital Humanities awarded the Reading the First Books project $215,000 to support an inter-institutional initiative to develop Ocular, a tool for the automatic transcription of multilingual, early modern printed books.[2] The grant also funded a collaboration with the Institute for Digital Humanities, Media, and Culture at Texas A&M University to build the DH Dashboard, a browser-based interface for automatically transcribing historical books.

In this paper, we reflect on the project's successes and challenges over the course of 2.5 years. Unlike the project report, here we are not focused on measurable outcomes. Instead, our discussion is oriented towards a public of teachers, practitioners, and future grant applicants; we are taking this opportunity to reflect on our project design and implementation. We are grateful to the members of that public who helped us to identify priorities in writing this document.[3]

# Part I: Project Highlights

Our most significant contributions include developing automatic transcription tools for early modern printed books, and developing an interface for automatic transcription.

## Multilingual OCR

Automatic transcription of early modern printed books is a challenge for multiple reasons. Early modern books have material qualities, like uneven inking and unfamiliar typefaces, that can make characters difficult to recognize (see Appendix B.1). They also have linguistic qualities that can make them hard to interpret, like variable orthography and radical multilinguality (see Appendix B.2, B.3).

Through the Reading the First Books project, we modified Ocular to handle the linguistic challenges of early modern transcription. Our modification of the tool automatically recognizes code-switching, even mid-sentence, and uses that information to more accurately transcribe the text. It also automatically learns orthographic variations, accurately identifying patterns such as the use of a "v" in place of a "u", or a "c" in place of a "z". In addition to allowing the tool to produce more accurate transcriptions, this feature has the additional benefit of enabling the joint production of diplomatic (historically accurate) and normalized (modernized) transcriptions. While the diplomatic transcription serves those interested in historical language use, the normalization feature allows for a more searchable corpus.

---

[2] For a more detailed discussion of the research, visit the project site, which has been preserved on archive.org: https://web.archive.org/web/20180215225629/sites.utexas.edu/firstbooks

[3] Special thanks to everyone who provided feedback and commentary on this document, including Linda Rodriguez, Regina Marie Mills, Jennifer Isasi, Ben Brumfield, and Twitter colleagues: @helsinHashtags, @Brett_Fujioka, @jasonrhody, @museums365, @ShawnaRoss, @benwbrum, @mlemweb, @JohnRosinbum, @LMRodriguez, @NorthCountrySOB, @emma_slayton, @johnruss28, @liblaurie, @leoba.

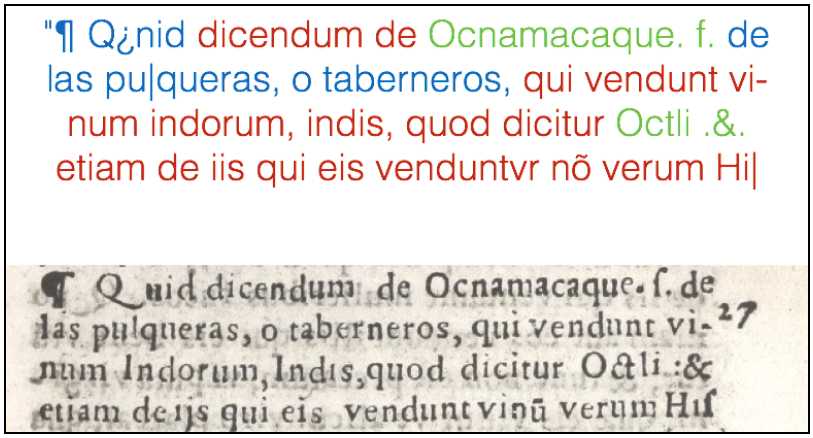


Figure 1: An automatically produced transcription of a radically multilingual text

## The DH Dashboard

Our original plan was to integrate Ocular into a pre-existing interface for automatic transcription, hosted by the Early Modern OCR Project (eMOP) at Texas A&M University.[4] It soon became clear that the infrastructure underlying that interface was too rigid for the integration of multiple tools. In response, eMOP made the decision to completely overhaul their system. Although this delayed our timeline, it turned out to be the correct decision. The overhauled DH Dashboard is a flexible system that enables transcription and evaluation across multiple tools. Ongoing development of the DH Dashboard, supported by a recent NHPRC-Mellon Digital Edition Publishing Cooperatives Grant, will enable the development of a robust REST API that will allow the DH Dashboard to serve as the backend for (or bridge between) an ecosystem of digital edition publishing tools.[5]

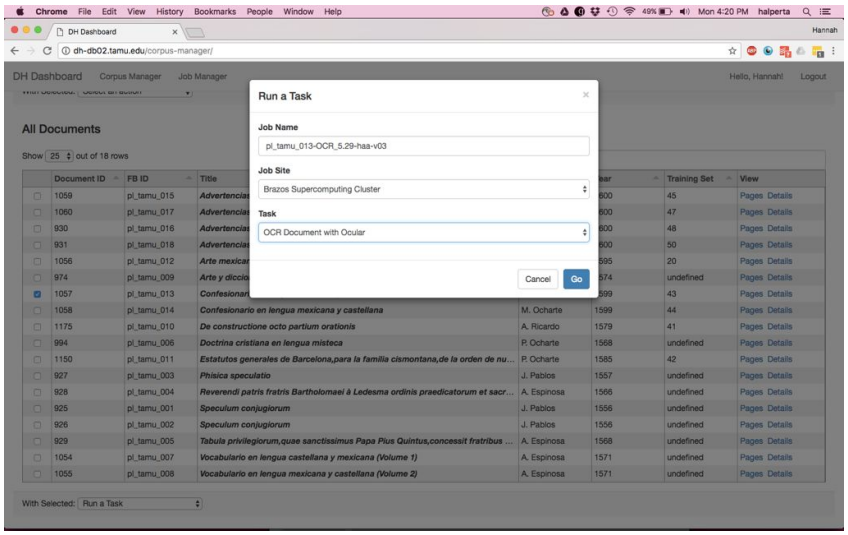


Figure 2: Screenshot of the DH Dashboard in action.

[4] http://emop.tamu.edu/
[5] https://libarts.tamu.edu/news/dr-laura-mandell-receives-grant-from-mellon-foundation/

# Part II: Project Proposal

The Reading the First Books project began as an attempt to create a text corpus for the *Primeros Libros*, a repository of over 300 digital facsimiles of books produced prior to 1601 in the Americas.[6] The books were available online as images, but could not be searched or analyzed computationally. Two graduate students, Hannah Alpert-Abrams and Dan Garrette, thought it would be trivial to transcribe those books automatically using Optical Character Recognition (OCR). The project was born out of the failures of those early transcription attempts.

## From Student Project to Institutional Initiative

The NEH proposal was drafted as the final assignment for an Introduction to Digital Humanities course at the UT Austin School of Information.[7] Expanding the project into a fully-fledged digital initiative required institutional investment at several levels. This included:

- **Project Management:** The library's Digital Scholarship Coordinator championed the project and oversaw the revision of the proposal and the creation of the project team.
- **Grant Writing:** The library's grants officer, Megan Scarborough, helped to write the budget and coordinated the relationship with the university's grants office.
- **Library Support:** A verbal and written commitment was acquired from IT and library leadership. That commitment included long-term preservation and web development.
- **Interinstitutional Partnerships:** A partnership was established with the Early Modern OCR Project at Texas A&M University.
- **Faculty Commitment:** A faculty member was invited to serve as the project PI.

Institutional support was vitally important both for developing this proposal and for conducting a project at this scale.

# Part III: Project Implementation

The Reading the First Books project began with the desire to transcribe the books in the Primeros Libros corpus, but it was motivated by the idea that OCR technology must address the multilingual and orthographically variable language which forms much of the documentary record. The project design reflects those dual priorities. It had four major stages:

1. **Development**: Establish a collaborative relationship between computational linguists and book historians that would enable research into improving systems for the automatic transcription of early modern books by developing Ocular, an experimental OCR tool.

[6] Learn more: http://primeroslibros.org/ . We are grateful to the PL partners for their support.

[7] The course was taught by Tanya Clement; the original assignment was to produce an abstract, statement of innovation, and statement of humanities significance for an 'imaginary' project according to the NEH Digital Humanities Start-Up Grant guidelines. Tanya Clement generously helped us by modifying the assignment for this project, which was better suited to an implementation grant. The syllabus is available online: INF 383H, Fall 2014.

2. **Integration**: Work collaboratively with the Early Modern OCR Project (eMOP) to build a user-friendly interface for the application of advanced OCR tools to early modern printed books.
3. **Implementation**: Use eMOP and Ocular to create a new transcribed corpus of printed books from early colonial New Spain.
4. **Documentation and Launch:** Host a symposium for project stakeholders, and document the project.

In our Project Report, we reflect on each of these elements in turn. For this document, we have chosen instead to consider three aspects of the project design and management: labor, methodology, and workflows.

## Labor

For the University of Texas project team, the bulk of the labor for this project was carried out by graduate students. This included: designing and implementing research; disseminating outcomes through writing and presentations; project management and documentation; and data creation and evaluation. The majority of this student labor was supported by the NEH in the form of a two-year Graduate Research Assistant (GRA) position, which was funded according to the standard University of Texas Library rates: $19,738 salary for a twelve-month appointment, plus $5,921 per year in fringe benefits and $9,665 per year in tuition assistance. Our research partners were supported through $3,000 per year in consultant fees. We are proud that we were able to provide substantial financial support for graduate student work through this project.

Two kinds of labor were undervalued in this project: project management and development. In the case of project management, our institution has shown support for this work through the 2011 creation of the Digital Scholarship Coordinator as a permanent staff position.[8] It was an unfortunate accident that the DS Coordinator retired just four months after the project began, and was not replaced until the following academic year. As a result, the work of project management was displaced onto the first GRA, who maintained the position even after departing from the project as a form of donated labor. These circumstances led us to a new appreciation of the labor of project management. They also had significant consequences for the project, because graduate students simply do not have enough institutional knowledge or power to handle the complex and often political negotiations involved in sustaining a digital project.

When we say we undervalued development, we refer specifically to the programming work required to maintain and troubleshoot the Ocular code. In the case of data management, web development, and preservation, the University of Texas and Texas A&M University both have paid staff responsible for this work; our grant supported their labor by paying part of their salaries. Our institutions did not have staff with the skills or mandate to help develop code, however. As a result, we leaned heavily on our computer scientist collaborators, but their availability was limited and their work was unpaid. Ultimately, our project was held back

[8] The Digital Scholarship Coordinator position was created in 2011 in response to the institutional restructuring that led to the merging of the Benson Latin American Collection, part of the University of Texas Libraries; and LLILAS Benson Latin American Studies, a program in the College of Liberal Arts. Prior to 2011, the position was based at the Benson.

significantly because we could not resolve bugs in the Ocular code. Perhaps funds directed towards a freelance programmer could have alleviated this problem.

An additional labor challenge that we faced was staff turnover. Of the twelve individuals named in the original proposal, only four remained in the same position throughout the course of the grant, and only seven remained involved in the project for its duration. We suspect this is not uncommon for academic projects. After a rocky start, we were able to weather these turnovers by prioritizing documentation, including meeting minutes and "letters of understanding" that recorded stakeholder commitments. Documentation was undertaken by our GRAs.

## Methods & Methodologies

Because they are so profoundly interdisciplinary, digital humanities methodologies can be difficult to define.

Our work developing tools for automatic transcription depended on methods from the fields of computational linguistics and media archaeology, described in full in the corresponding publications (see Appendix A). In the case of computational linguistics, we developed an experimental design that allowed for statistically valid evaluations of transcription error rates; this approach seems to be the standard in the field. We paired this with a close-reading of the evaluation that allowed us to hypothesize about where our system was successful, and where there was room for further improvement. In the case of media archaeology, we paired close reading with historical analysis to offer a situated discussion of digital technology in cultural context. We found it was most effective to apply these methods independently when writing for various venues. In public presentations, however, we benefited from a more fluid movement between the two kinds of analysis; we all developed some fluency in speaking across disciplines.

The methods underscoring our integration work were more ad-hoc. In collaborating with eMOP to develop the DH Dashboard, a digital interface for OCR, we worked iteratively to design and test solutions, depending primarily on experience-based evaluation. We were the first in a series of user groups to test the functionality of the DH Dashboard , and contributed to an iterative feedback process that improved the system's design and user experience.

Over the course of this project, we became increasingly interested in applying a decolonial critique to our methodologies. The Primeros Libros project, on which our project was built, is a post-custodial collection, an effort to collect and disseminate historical materials without replicating a colonialist approach to cultural heritage.[9] To what extent did our project, and its methodologies, support that same mission?

At a superficial level, our focus on improving access and discovery of colonial materials feels like a move in the right direction. The materials that we were working with, sixteenth-century Mexican books, have long been kept apart from the colonial and indigenous communities that produced them. Libraries, in particular, have served to isolate texts from affiliated communities. The *Primeros Libros* project uses the language of "digital repatriation" to describe the ways that it works to counteract that colonial past. By facilitating access to that collection, we hoped to be part of the work of undoing this colonial history. Furthermore, by

[9] The Society of American Archivists defines the post-custodial model as "the idea that archivists will no longer physically acquire and maintain records, but that they will provide management oversight for records that will remain in the custody of the record creators" (archivists.org). See Christian Kelleher, "Archives Without Archives." Journal of Critical Library and Information Studies 1.2 (2017). https://doi.org/10.24242/jclis.v1i2.29

bringing attention to the ways that OCR tools exclude culturally complex textual artifacts, we hoped to counteract the monolingualism and anglocentrism of technical research.

Decolonizing methodology is not so simple, however. As Linda Tuhiwai Smith writes, research is "embedded in imperialist expansionism and colonization."[10] We found that the colonial documents we were working with actively resisted the methods that we applied to them. When evaluating transcription accuracy, for example, we struggled with the limitations of statistical measures to account for the slippages of indigenous inscription, especially in a context where orthographies were variable and inscribed language was an imperfect representation of a resistant voice. There is no evidence, furthermore, that transcribing these documents fulfills any need or priority put forth by Mexican or indigenous communities. These are the priorities of our institutions, and our research community, and we acknowledge that here.

We struggled, too, with the way that our project design concentrated resources within the academy, rather than helping to distribute them to affiliated communities. In future resources, we hope to correct for that weakness in our project design.

# Workflows

## Stage 1: Development

Plan: Fall 2015 (4 months)
Reality: Fall 2015 (5 months)

### People

- Kent Norsworthy (project management) (3 hrs/week)[11]
- Hannah Alpert-Abrams and Dan Garrette (Ocular R&D) (20hrs/week)
- Stephanie Wood (language consultation) (10 hours)

### Process

This stage included developing Ocular and gathering language data. We are lucky that the research took about as long as we expected.

## Stage 2: Integration

Plan: Winter 2016 - Summer 2016 (8 months)
Reality: Winter 2016 - Spring 2017 (16 months)

### People

- Hannah Alpert-Abrams (project management) (20hrs/week, 4 mos; 5 hrs/week, 12 mos)
- Trey Dockendorf, Bryan Tarpley, Matt Christy (dashboard development) (20 hrs/week)
- Maria Victoria Fernandez (dashboard testing) (20hrs/week, 12 mos)

[10] Linda Tuhiwai Smith. *Decolonizing Methodologies: Research and Indigenous Peoples.* Zed Books: London, 2012.
[11] We did not keep track of hours, so these are very rough estimates. We wanted to give a sense of how work was distributed and who was involved. It should be noted that there were additional people involved at a lower rate.

- Dan Garrette (Ocular support) (5 hrs/week)

### Process

This stage, which involved integrating Ocular with the DH Dashboard, took longer than we foresaw for two reasons. First, the two developers responsible for the project left their positions. We attempted to increase the pace of the project to complete it before their departure, but this only led to more confusion and delays. When the new developer was hired some months later, we were forced to start again from scratch.

Thanks largely to his brilliance and strategic vision, we ended up realizing that with more time, the project could accomplish a lot more than we had originally hoped. The redesign took a lot more time than we expected, but we are pleased with the outcome, which we hope will have significant impact on automatic transcription workflows in the future.

During this time, we also worked with metadata librarians at the University of Texas Libraries to develop metadata standards for Ocular's xml output that would meet the needs of our web developers so that we could make the material available to the public.

## Stage 3: Implementation

Plan: Fall 2016 - Spring 2017 (8 months)
Reality: Summer 2017 - Spring 2018 (ongoing)

### People

- Hannah Alpert-Abrams (project management) (5-10 hrs/week)
- Maria Victoria Fernandez (GRA) (20 hrs/week, later 5hrs/week)
- Benn Chang (data preservation (15 hours)
- Melanie Cofield (metadata) (20 hours)
- Audrey Templeton (web development) (unknown)
- Dan Garrette (Ocular support) (40 hours)

### Process

The goal of this stage was to transcribe the 300+ books in the Primeros Libros corpus. We faced multiple challenges at this point. The delay in the DH Dashboard meant we started later than we originally planned. Furthermore, limitations to and bugs in the Ocular code kept us from transcribing on schedule (and ultimately forced us to reevaluate our goals for this stage).

We were able to receive a 'no-cost extension' from the NEH for the completion of this work, but without additional funds, we had limited capacity to continue. The GRA who led the second year of work was assigned to other projects. No other students or staff members at UT Austin had a mandate to continue the work. While we did transcribe 50 of the books in the corpus, we largely consider this phase permanently incomplete. We are open to continuing the project but have no concrete plans to do so at this time.

A second goal of this stage was to update the Primeros Libros project website to incorporate the transcriptions. Due to capacity challenges on the part of our development partners, this work is still under development.

### Stage 4: Documentation and Launch

Plan: Summer 2017 (3 months)
Reality: Summer 2017 - March 2018 (9 months)

#### People

- Maria Victoria Fernandez (5 hrs/week)
- Hannah Alpert-Abrams (5 hrs/week)
- Albert Palacios (symposium planning) (120 hours)

We chose to launch the project as planned, with a very successful one-day symposium. This allowed us to acknowledge and celebrate the successes of the project even as we worked to resolve ongoing challenges.

Thanks to a no-cost extension, we were able to use some dedicated GRA time to document the project during the fall. This ended in December 2017, with the departure of the GRA. No staff time was dedicated to the completion of the project report or white paper in 2018.

# Part IV: Project Outcomes

## Collaboration & Publication

Digital humanities projects can be difficult to disseminate and often have designs that erase labor. We decided to publish across platforms and disciplines, a practice that sometimes felt as though it doubled our work, but that had the impact of disseminating our work quite broadly. This included peer reviewed work and public presentations both in computational linguistics and digital humanities, as well as publications in more popular non-reviewed publications such as our project blog. Through this method, we sought to highlight the voice of multiple contributors to the project in ways that would directly support their professional aspirations and learning needs.

We are particularly proud of our final symposium, which brought together scholars, librarians, developers, and students to discuss the project's accomplishments and challenges, the future of colonial digital materials, and how digital scholarship will facilitate further engagement with colonial Latin American history. A central goal of the symposium was to highlight the collaborative work of librarians, developers, and researchers involved in the First Books project. Three roundtables highlighted, respectively, the multi-institutional and interdisciplinary nature of the project; digital scholarship in colonial Latin American studies; and the future of digital scholarship as viewed by graduate students. We are particularly proud of the diversity of these roundtables, which included tenured faculty, library staff, administrative leaders, and graduate students from multiple institutions. Nearly seventy individuals took part in the symposium, representing thirteen departments, the academic libraries, and several local software developers.

# Assets for the Academic Commons

## Publications

All of our peer-reviewed publications (see Appendix A) were published through open access venues and are available online. Our website, including our blog, is also archived via archive.org.

## Tools

Ocular, the tool for the automatic transcription of early modern printed books, is open source and available via GitHub.[12] Our contributions to Ocular include the ability to handle multilingual documents, the ability to handle multiple orthographies, and the ability to jointly transcribe diplomatic (historical) and normalized (modern) versions of the text.

The DH Dashboard, the eMOP interface for using Ocular and other OCR tools, remains in the development phase. The project has expanded significantly thanks to a recent NHPRC-Mellon Digital Edition Publishing Cooperatives Grant.[13] Currently, there is a beta testing program available for researchers, [14], with plans to eventually make a hosted version of the tool available for public use. The project is also open source.[15]

## Data Sets

The Primeros Libros corpus has more than 300 digital facsimiles of books printed prior to 1601 in the Americas. We aspired to transcribe that collection, which contains text in eight languages, in its entirety. However, we were only able to transcribe 50 of those books using Ocular and the DH Dashboard, in only four languages (Spanish, Latin, Nahuatl, and Zapotec).

The OCR for those pages can be viewed and downloaded through a temporary staging site; the University of Texas Libraries has not yet decided whether to incorporate these transcriptions into the project website.[16]

Because of their poor quality, and because of the delays in the production of these texts, we did not conduct a statistical evaluation of transcription quality on these texts. We are working to describe their quality more fully for future use.

---

[12] https://github.com/tberg12/ocular
[13] https://libarts.tamu.edu/news/dr-laura-mandell-receives-grant-from-mellon-foundation/
[14] For more information, contact Bryan Tarpley: bptarpley@tamu.edu
[15] https://gitlab.dh.tamu.edu/bptarpley/dh_dashboard
https://gitlab.dh.tamu.edu/bptarpley/dh_corpus
https://gitlab.dh.tamu.edu/bptarpley/dh_job
https://gitlab.dh.tamu.edu/bptarpley/dh_agent
[16] halperta.com/firstbooks

# Part V: Future Research

## The Public Good

Several reviewers have asked us to describe the ways that this project benefits the public good, including academia, secondary school education, and the general public.

We cannot claim to have this level of impact on our world. However, we do believe that we have influenced smaller corners of our community in meaningful ways.

## Optical Character Recognition

When we began this project, work to develop automatic transcription tools in the United States was focused primarily on monolingual and Anglophone texts. While we are not the only ones who advocated for a reorientation towards orthographic and linguistic complexity, we do believe that we have helped to set the groundwork for continued research on these historical documents. We see evidence of this in projects such as the Historical and Multilingual OCR workshop, which was held at Northeastern University in February 2018, and which several project participants attended.[17] The workshop brings together humanists, researchers, and industry specialists interested in OCR to identify current needs in Optical Character Recognition and to chart a path forward for the development of OCR software. While the challenges addressed by this project are myriad, we believe that this work has contributed to that conversation and been influential in the ongoing development of OCR software. We also believe that the DH Dashboard will be fundamental to the future of this work. The conference has left us optimistic about the future of OCR for low-resource languages and for historical books.

## Digital Humanities and Colonial Latin American Studies

Digital humanities has long been supported in the field of Latin American Studies by groups like the Red de Humanidades Digitales and Global Outlook::Digital Humanities.[18] But discussions around colonial DH remain nascent. We hope that the community established through the First Books symposium, in addition to connections made through conference presentations, will help to strengthen this field, attract new practitioners, and lend legitimacy (and resources) to those already working to reimagine colonial studies through a digital lens.

## Digital Humanities at the University of Texas at Austin

Digital humanities at UT Austin has changed substantially over the course of the Reading the First Books project. Originally a standalone project, the work was soon incorporated into a newly formed Digital Initiatives team led by Albert A. Palacios, Theresa Polk, and Susan Kung, and housed at LLILAS Benson. This group has since launched a series of initiatives related to digitizing, hosting, and providing access to vulnerable Latin American documents. In addition, the group has begun to develop multilingual corpora for more advanced cultural analytics of

[17] https://ocr.northeastern.edu/

[18] http://www.humanidadesdigitales.net/acerca-de/ and http://www.globaloutlookdh.org/

Latin American materials. Though we were not able to produce a significant contribution to these collections-as-data, the project did set a framework for conducting large-scale digital humanities research initiatives within Latin American Studies at UT Austin. The Digital Initiatives in Latin American Studies team has in turn served as a model for digital initiatives across the university system more broadly.

## Future Research

There are so many areas for future research that we believe this project has opened! Here are some of the directions we would like to see this work go:

- **Page segmentation**: How do you know where one line, or block of text, or image, begins and the previous one ends? The weakness of the page segmentation tool that Ocular uses was one of the major challenges that we faced, but from a research standpoint, it's not the most exciting one. The real challenge has to do with sequence: how do you not merely segment the pieces on the page, but categorize them according to reading flow? We have heard this problem defined most succinctly as: how do you automatically segment a page so that it can be processed properly by a screen reader?

- **Normalization:** In our work, we began to explore the possibility of jointly outputting diplomatic and normalized versions of a text: one transcription that preserves the historical spelling, and another one that reflects modern standards. We found that this produces more accurate transcriptions; it also improves downstream accuracy for tasks like searching or parsing. But we only scratched the surface of this challenge, and the accuracy of our normalized transcription remains low. There is so much more research to be done to improve this transcription feature.

- **OCR as an analytic tool:** What if OCR was not just a transcription system, but also a tool that analyzed the properties of a written page? In collaboration with a group at Carnegie Mellon University, we have been pushing this question to learn new things about book history.[19] But there is so much more to learn about how the statistical analysis of a historical page can provide insight into the qualities of the text itself.

- **OCR Evaluation:** How good is your OCR output, and how good do you want it to be? There is a real need for standardized approaches to OCR evaluation, approaches that can be applied broadly as a precursor to digital scholarship conducted using OCR text. This project is bigger than it first appears, and if we had time and resources, it's where we would go next.

- **Decolonial DH:** In our discussion, we consider some of the ways that colonial and imperial contexts impacted our project. But this is just the beginning, and we would love to see projects take this further to help lay the groundwork for the ethical digitization and digital exploration of colonial materials.

[19] PI: Taylor Berg-Kirkpatrick.

# Appendices

## Appendix A: Publications

### Articles

Garrette, Dan, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. "Unsupervised Code-Switching for Multilingual Historical Document Transcription." *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Denver, CO: 2015.

Garrette, Dan, and Hannah Alpert-Abrams. "An Unsupervised Model of Orthographic Variation for Historical Document Transcription." In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*: San Diego, CA: 2016.

Alpert-Abrams, Hannah. "Machine Reading the Primeros Libros." *Digital Humanities Quarterly* 10, no. 4 (2016). http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html.

Alpert-Abrams, Hannah, and Maria Victoria Fernandez. "Reading the First Books: Colonial Mexican Documents in the Digital Age." *Portal: LLILAS Benson Latin American Studies and Collections Annual Review*, Austin, TX: 2017.

### Presentations

Alpert-Abrams, Hannah, and Dan Garrette. "Automatic Transcription in Colonial Contexts." Poster presented at the Texas Digital Humanities Consortium (TXDHC) Conference, University of Texas at Arlington, April 9-11, 2015.

Alpert-Abrams, Hannah. "The Electronic Edition of Colonial and Nineteenth-Century Latin American Texts: New Tools, New Models for Collaboration." Workshop panel at the XXXIV International Congress of the Latin American Studies Association, New York, NY, May 27-30, 2016.

Fernandez, Maria Victoria. "Reading the First Books: Enhancing Digital Archives through Automatic Transcription of Early Modern Printed Documents" Presentation at the Society of American Archivists Conference, Atlanta, GA, July 31-August 6, 2016.

Alpert-Abrams, Hannah. "Esta ãphibologia no ay ē latin: Machine reading linguistic hybridity in the Primeros Libros." Presentation at the Tercer Encuentro de Humanidades Digitales, Mexico City, Mexico, September 12-14, 2016.

Alpert-Abrams, Hannah. "Reading the First Books" Presentation at the John Carter Brown Library, Providence, RI, November 2016.

Alpert-Abrams, Hannah. "Optical Character Recognition." Presentation at the Modern Language Association Annual Convention, Philadelphia, PA, January 5-8, 2017.

Alpert-Abrams, Hannah. "Transcribing Historical Texts." Presentation and tutorial at the John Carter Brown Library, Providence, RI, March 2017.

Fernandez, Maria Victoria. "Managing and Preserving Data in a Cross-Institutional Digital Humanities Initiative: Lessons from the Reading the First Books Project." Presentation at the UT Digital Preservation Symposium, University of Texas Libraries, Austin, TX, April 21, 2017.

Alpert-Abrams, Hannah. "Reading the First Books Project Introduction." Presentation at the "Reading the First Books: Colonial Documents in the Digital Age Symposium," University of Texas at Austin, May 30, 2017.

Tarpley, Bryan. "Leveraging eMOP Project Assets in the Service of Reading los Primeros Libros" Presentation at the "Reading the First Books: Colonial Documents in the Digital Age Symposium," University of Texas at Austin, May 30, 2017.

Fernandez, Maria Victoria "Breakdowns in Machine Reading: Confronting the Constructive Limitations of Decolonializing DH." Presented at the Digital Frontiers Conference, University of North Texas, Denton, TX, September 21-23, 2017.

Tarpley, Bryan. "Breakdowns in Machine Reading: Attempting to De-privilege Modern English Print with the Power of Supercomputing and the DH Dashboard." Presented at the Digital Frontiers Conference, University of North Texas, Denton, TX, September 21-23, 2017.

## News Coverage

"NEH Grant Will Transform Study of Early Books." *Benson Latin American Collection, University of Texas Libraries*. July 31, 2015. https://www.lib.utexas.edu/benson/announcements/neh-grant-will-transform-study-early-books

"NEH Grant Will Transform Study of Early Books." *Teresa Lozano Long Institute of Latin American Studies.* July 31, 2015. https://liberalarts.utexas.edu/llilas/news/article.php?id=9649

"NEH Grant Will Transform Study of Early Books." *UT News.* August 4, 2015. https://news.utexas.edu/2015/08/04/neh-grant-will-transform-study-of-early-books

Garcia, Carlos. "UT Researchers Using Computer Program to Decipher Early Latin American Text." *Spectrum News*. Austin, TX. May 30, 2017. http://spectrumlocalnews.com/tx/austin/top-stories/2017/05/30/computer-program-deciphers-early-latin-american-text

Alpert-Abrams, Hannah. "Reading the First Books: Annual Report from Hannah Alpert-Abrams" *Program in Comparative Literature, The University of Texas at Austin.* February 8, 2017. https://liberalarts.utexas.edu/complit/news/11215

## Appendix B: Figures

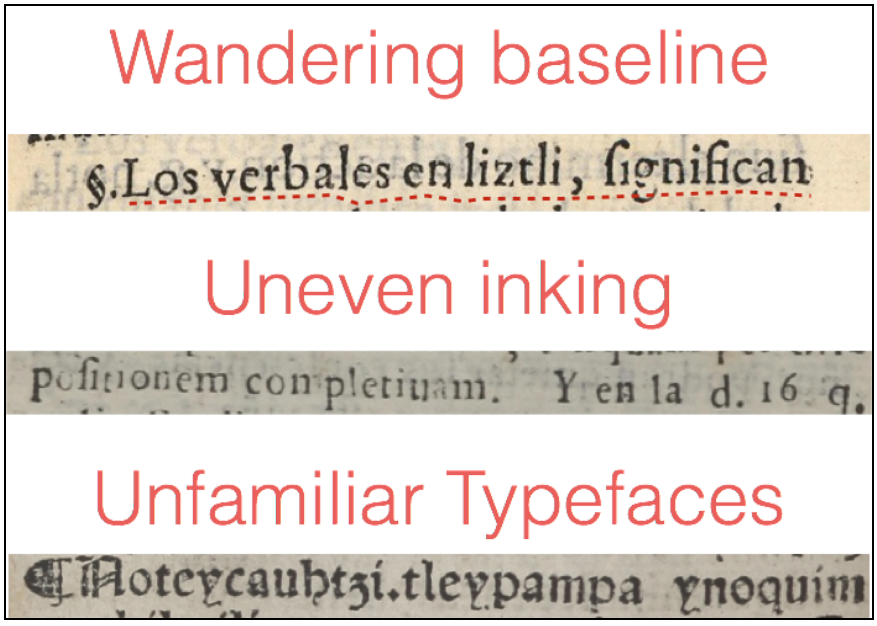


Figure 1: Material challenges of early modern printing.

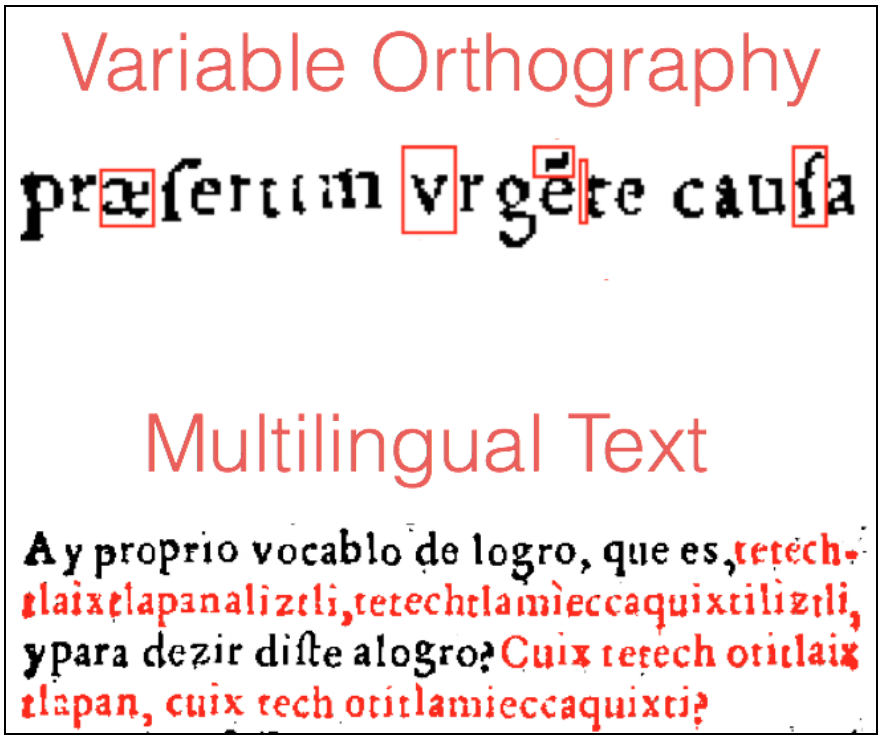


Figure 2: Linguistic challenges of early modern printing.

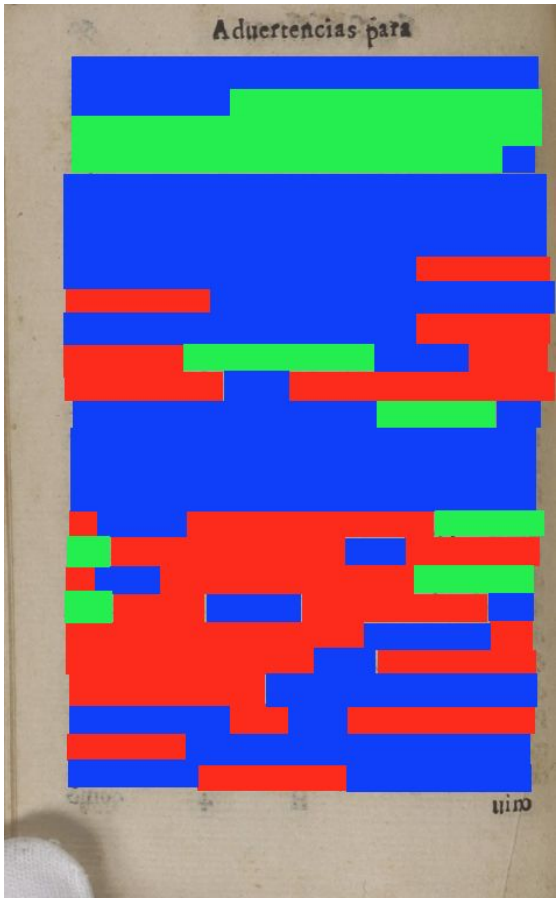


Figure 3: Radical multilinguality in the *Primeros Libros* corpus (red is Spanish, blue is Latin, and red  is Nahuatl).

# Appendix C: People

## Project Managers:

- Albert Palacios, Latin American Studies Digital Scholarship Coordinator, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Theresa Polk, Post-Custodial Archivist, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Kent Norsworthy, (former) Digital Scholarship Coordinator, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin

## Graduate Research Assistants:

- Maria Victoria Fernandez, Digital Scholarship Graduate Research Assistant, LLILAS Benson Latin American Studies and Collections, and Master's student in Information Studies and Latin American Studies.
- Hannah Alpert-Abrams, PhD candidate in the Program in Comparative Literature, The University of Texas at Austin

## Project Director:

- Sergio Romero, Assistant Professor, Teresa Lozano Long Institute of Latin American Studies and Department of Spanish & Portuguese, The University of Texas at Austin

## eMOP Implementation Team

- Liz Grumbach, Research Associate, Initiative for Digital Humanities, Media, and Culture, and Project Manager of the Advanced Research Consortium (ARC) and 18thConnect, Texas A&M University
- Bryan Tarpley, Lead Software Applications Developer, Initiative for Digital Humanities, Media, and Culture, Texas A&M University.
- Project Co-Director: Laura Mandell, Director, Initiative for Digital Humanities, Media, and Culture, and Professor, Department of English, Texas A&M University.
- Project Co-Director: Anton duPlessis, Colonial Mexican Collection Curator, Cushing Memorial Library, Texas A&M University
- Matt Christy, (former) Lead Software Applications Developer, Department of English, Texas A&M University
- Trey Dockendorf, (former) Systems Analyst I, The Academy for Advanced Telecommunications and Learning Technology

## Outside Consultants

- Stephanie Wood, Director and Senior Research Associate, Wired Humanities Projects, University of Oregon
- Dan Garrette, Postdoctoral Fellow, University of Washington

## Metadata, Preservation, and Web Development

- Ting-Benn Chang, Technology Coordinator for Digital Stewardship, University of Texas Libraries, The University of Texas at Austin
- Melanie Cofield, Metadata Coordinator, University of Texas Libraries, The University of Texas at Austin
- Frederick Gilmore, DevOps Lead & Senior Systems Administrator, University of Texas Libraries, The University of Texas at Austin
- Audrey Templeton, Software Developer/Analyst, University of Texas Libraries, The University of Texas at Austin

## Advisory Board

- Tanya Clement, Assistant Professor, School of Information, The University of Texas at Austin
- Matt Cohen, Associate Professor, Dept. of English, The University of Texas at Austin
- Adam Coon, Assistant Professor, Dept. of Spanish, The University of Minnesota, Morris.
- Susan Deans-Smith, Associate Professor, Colonial Latin American History, Dept. of History, University of Texas at Austin
- Julianne Gilland, Associate Director for Scholarly Resources and Special Collections Curator, Benson Latin American Collection, The University of Texas at Austin
- Kelly McDonough, Assistant Professor, Dept. of Spanish and Portuguese, The University of Texas at Austin
- Albert Palacios, Doctoral Candidate, Dept. of History, The University of Texas at Austin
- Kenneth C. Ward, Maury A. Bromsen Curator of Latin American Books, The John Carter Brown Library, Brown University
- Nicholas Woodward, Information Technology Specialist, Office of Strategic Initiatives, The Library of Congress

## Other Support

- Virginia Burnett, Director, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Charles Hale, (former) Director, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Heather Gatlin, Executive Director, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Megan Scarborough, Grant Writer, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Susanna Sharpe, Communications Coordinator, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin
- Ruth Sogas-Paramio, LLILAS Benson Latin American Studies and Collections, The University of Texas at Austin